IS 6733: Deep Learning
Homework 1

P1 (50 pt): Write a Python code in Colab using Pandas and/or Matplotlib libraries to accomplish the following tasks:

1. Import the iris flowers dataset using pandas.read_csv() with the following URL link (10pt); The DataFrame must have the following column names: 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)', and 'class'; (5pt) Print the first 5 rows of the DataFrame (5pt).
   a. Dataset source file http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
   b. Dataset description: http://archive.ics.uci.edu/ml/datasets/iris
   c. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html
      - You can fetch the data online by inputting the above URL in pandas.read_csv(url = XXX). Downloading the data to a local copy will make the shared Colab code in your homework submission inexecutable.
      - Pay attention to the header and index_col arguments when using read_csv().
2. Summarize the dataset
   a. Print out a concise summary of the DataFrame using .info() and the shape of the DataFrame (5 pt)
   b. Print out the statistics of the continuous columns using .describe() (i.e., the four attribute columns) (5 pt)
   c. Print out the number of rows that belong to each class (5 pt)
3. Data visualization
   a. Separate out the first four columns of the original DataFrame into a new DataFrame and print out the first 5 rows of the new DataFrame (5 pt)
   b. Univariate Plots: plot a histogram for each column of the new DataFrame (5 pt)
   c. Multivariate Plots: plot a scatter plot for each pair of the columns of the new DataFrame using the *pandas.plotting.scatter_matrix* function (5 pt)
      - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.plotting.scatter_matrix.html

P2 (50 pt): Write a Python code in Colab using Pandas library to accomplish the following tasks:

1. Import the *Census Income (Adult)* dataset using Pandas, use the 14 attribute names (i.e., "age", "workclass", ….., "native-country") as explained in the dataset description as the first 14 column names and "salary" as the last column name (5 pt) , view the strings '?', ' ?', '? ', or ' ? ' as the missing values and replace them with NaN (the default missing value marker in Pandas) (10 pt), and print out the first five rows of the DataFrame. (5 pt)
   a. Dataset source file: http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
   b. Dataset description: http://archive.ics.uci.edu/ml/datasets/census+income
   c. Pay attention to the header and index_col arguments when using pandas.read_csv().

2. Dataset checking and cleaning
    a. Print out a concise summary of the DataFrame and observe if null values exist in each column of the DataFrame by checking the summary (10 pt)
    b. Filter out the rows that contain missing values and print them out (10 pt)
    c. Drop the rows of the DataFrame with missing values using .dropna() and observe if null values still exist in each column by checking the concise summary again (10 pt)

Submission Instruction: Submit a PDF file of your codes and outputs and a public Google Colab shared link to your source file (.ipynb format) to Blackboard (See the submission details on Blackboard).